Evolution and RNA Relics. A Systems Biology View

Jacques Demongeot, Nicolas Glade, Andres Moreira

Evolution and RNA Relics. A Systems Biology View

Jacques Demongeot · Nicolas Glade · Andrés Moreira

Abstract The genetic code has evolved from its initial non-degenerate wobble version until reaching its present state of degeneracy. By using the stereochemical hypothesis, we revisit the problem of codon assignations to the synonymy classes of amino-acids. We obtain these classes with a simple classifier based on physico-chemical properties of nucleic bases, like hydrophobicity and molecular weight. Then we propose simple RNA (or more generally XNA, with X for D, P or R) ring structures that present, overlap included, one and only one codon by synonymy class as solutions of a combinatory variational problem. We compare these solutions to sequences of present RNAs considered as relics, with a high interspecific invariance, like invariant parts of ^{t}RNAs and micro-RNAs. We conclude by emphasizing some optimal properties of the genetic code.

Keywords Genetic code · Primitive RNA · Archetypal RNA ring · RNA relics · Micro-RNA · siRNA

1 Introduction

J. Demongeot (✉) · N. Glade · A. Moreira
TIMC-IMAG, UMR CNRS 5525, Faculty of Medicine of Grenoble, University J. Fourier, 38 700 La Tronche, France
e-mail: Jacques.Demongeot@imag.fr

J. Demongeot
Institut Universitaire de France, Paris, France

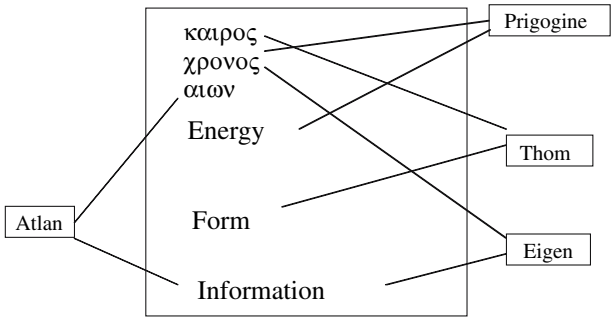


Fig. 1 The "1972" paradigm

Understanding the first steps of the assembly of the primitive molecules of life involves the existence of an initial replicable structure, which allows the start and the progressive complexification of living systems (Atlan 2004).

Twenty years after the numerous scientific discoveries of the year 1953 (DNA by Watson and Crick, Miller's experiment about the origin of life, cellular automata by von Neumann, free monoids by Schützenberger and self-organization by von Foerster), the "1972" paradigm (Fig. 1) comes from a deep reflexion about the role of the time in biology, time being taken in all acceptations defined by the Greeks: *καιρος* or convenient time of Herophilus (birth time of the archetypal forms, for Heraclites and Thom), *χρονος* of Hippocrates (measuring the dissipation of energy, for Prigogine and Eigen), and *αιων* or immanent time, the substrate of the Aristoteles psyche, time without quantum (with only order) in which information exchanges occur, for Atlan (Glansdorff and Prigogine 1971; Eigen 1971; Thom 1972; Atlan 1972). Thirty-six years after the foundation of this paradigm, disciplines like systems biology and those that study the complexity of the life revisit it for explaining the most recent biological data. Such approaches use the main concepts emphasized in 1972: (i) the notion of biological time at two levels, both in a global evolutionary perspective and also in a local energetic context, allowing to obtain optimal (i.e. satisfying a variational criterion) molecular forms or structures and (ii) the notion of biological information showing that these structures permit the conservation of ancestral functions (e.g. peptide building) by memorizing invariant intra- and interspecific molecular sequences. In this mind, we will describe in the following a mathematically plausible primitive genome consistent with biological facts, thus giving internal coherence to the stereochemical theory of the genetic code.

2 Origins of Life

2.1 Archetypal Genome: A Plausible Assembly

Born about 3.8 billions of years ago, modern forms of life started by the self-assembly of primitive molecules, e.g. Ribo-Nucleic Acids (RNAs) meeting Amino-Acids (AAs) in confining media like argils as the Montmorillonite (first proposed by A. Katchalsky)

or Oparine's coacervates (Oparine 1924): these self-assemblages could produce hybrid structures linking with weak bindings (electrostatic or van der Waals bonds) AAs with RNAs, e.g. AAs with the triplets of their synonymy class. After this transient contact, AAs would be strongly linked (by peptidic bonds) giving the first peptides, which then catalyze RNA synthesis in a positive loop of interactions. RNAs can be more generally replaced by XNAs (X for D, P or R) by emphasizing that there are other possible structures whose articulated skeleton brings together nucleic bases, helping the AAs confinement (a kind of "matrimonial agency" favouring marriages between nucleic sequences and AAs).

One of the most productive views about the first steps of life linking RNAs and AAs is the stereochemical theory of the genetic code (Pelc and Welton 1966; Welton and Pelc 1966; Hendry et al. 1981). This theory claims that triplets of the code are closely associated to amino-acids in a unique and degenerate way through a physico-chemical similarity causing their affinity and defined from at least three features: the Hydrophobicity H, the molecular weight P and the ability to Link L equal to the maximum of possible hydrogen bonds divided by the length of the maximal carbon chain inside the molecule (Blalock and Bost 1986; Mitaku et al. 2002). The correspondence between coding triplets and AAs is summarized on the Fig. 2: variables H, P and L are valued for each AA showing a structured organization of triplets into synonymy classes of sizes from 1 to 6. Following the wobble hypothesis by Crick (assuming the existence of an initial non degenerate coding of only 16 AAs by 16 pairs of nucleic bases), the two first bases of a triplet are decisive for its assignation to a synonymy class: the second base allows to roughly divide triplets space into classes associated to hydrophobic AAs if the

	U	C	A	G	
U	UUU Phénylalanine 2.8 3/7,165	UCU Sérine H=-0.8 L=4/3,P=105	UAU Tyrosine -1.3 4/7,181	UGU Cystéine 2.5 3/3, 121	U
	UUC Phénylalanine	UCC Sérine	UAC Tyrosine	UGC Cystéine	C
	UUA Leucine 3.8 3/5,131	UCA Sérine	UAA Stop	UGA Stop	A
	UUG Leucine P>110	UCG Sérine P<120	UAG Stop P>120	UGG Tryptophane -0.9 4/8,204	G
C	CUU Leucine 3.8 3/5,131	CCU Proline -1.6 3/4,115	CAU Histidine -3.2 5/6,155	CGU Arginine -4.5 6/6,174	U
	CUC Leucine	CCC Proline	CAC Histidine	CGC Arginine	C
	CUA Leucine H>0,	CCA Proline L≤3/3 H<0,	CAA Glutamine -3.5 L≥3/4 5/6,147	CGA Arginine	A
	CUG Leucine	CCG Proline	CAG Glutamine	CGG Arginine	G
A	AUU Isoleucine 4.5 3/5,131	ACU Thréonine -0.7 4/4,119	AAU Asparagine -3.5 5/5,132	AGU Sérine -0.8 4/3, 105	U
	AUC Isoleucine	ACC Thréonine	AAC Asparagine	AGC Sérine	C
	AUA Isoleucine	ACA Thréonine	AAA Lysine -3.9 4/6,146	AGA Arginine -4.5 6/6,174	A
	AUG Méthionine/Start 1.9 3/5,149	ACG Thréonine	AAG Lysine	AGG Arginine P<110	G
G	GUU Valine 4.2 3/4,117	GCU Alanine 1.8 3/3,89	GAU Aspartate -3.5 5/4,133	GGU Glycine -0.4 3/3, 75	U
	GUC Valine	GCC Alanine	GAC Aspartate	GGC Glycine	C
	GUA Valine	GCA Alanine	GAA Glutamate -3.5 5/5,147	GGA Glycine	A
	GUG Valine	GCG Alanine	GAG Glutamate	GGG Glycine	G

Fig. 2 Genetic code with hydrophobicity H, molecular weight P and binding capacity L

central base B is Uridin U or Cytosin C, heavy AAs if B is U or Adenin A, or AAs capable of hydrogen binding if B is C or Guanine G (Pelc and Welton 1966). Each nucleic base will be represented by a binary number of three bits, e.g. 110 for U: 1 for hydrophobicity, 1 for heaviness and 0 for hydrogen binding ability. In the following for the sake of simplicity we will cancel the third bit because coding with two bits suffices to explain the essential of the degeneracy (according to Pelc and Welton 1966). In the context of the direct stereochemical interaction between the AAs and their codons, the third base only binds with the head of AAs (the $NH3^{+}$–CαH–COO^{-} part) except for glycine and alanine, the smallest AAs. It accounts for the genetic code optimality (Gilis et al. 2001) and is classical for nucleic bases (Demongeot and Besson 1983): U = 11, C = 10, A = 01, G = 00.

The stereochemical theory has been proposed for explaining the degeneracy of the present genetic code. By implementing it in an automaton, we obtain a classification of synonymic codons close to the real one. This automaton called SUSY for «SUrrogate SYstem» (Demongeot et al. 2006) plays the role of a classifier associating triplets and AAs having same physico-chemical characteristics, like hydrophobicity $H \geq 0$ (resp. hydrophilicity $H < 0$) and high (resp. small) molecular weight P (positively correlated to L). SUSY acts by definition through a transition T on the set of binary numbers of six bits representing triplets:

$$\begin{aligned} T(t+1) &= \mathrm{MinV}(T(t)), \text{ with } V(T(t)) \\ &= \{X;\ S(X,T(t)) \geq 10 \text{ and } D(X,T(t)) \leq 3/32\}, \end{aligned}$$

where MinV is the minimum of the set V for the wobble order, D is a weighted Hamming distance $D(X,Y) = |x_3 - y_3| + |x_4 - y_4|/2 + |x_1 - y_1|/4 + |x_2 - y_2|/8 + |x_5 - y_5|/16 + |x_6 - y_6|/32$, S is a similarity criterion $S(X,Y) = \sum_{i=1,5} a_i(x_i,y_i)x_iy_i + b_i(x_i,y_i)(1-x_i)(1-y_i)$, where coefficients a_i and b_i are fixed by electrostatic (hydrophobicity H) and steric (molecular weight P) properties of the AAs (Trinquier and Sanejouand 1998):

$$a_1(1,1) = 2,\ a_2(1,1) = 2,\ a_3(1,1) = 3,\ a_4(1,1) = 2,\ a_5(1,1) = 2,\ b_1(0,0) = 2,$$
$$b_2(0,0) = 3,\ b_3(0,0) = 2,\ b_4(0,0) = 3,\ b_5(0,0) = 3.$$

The threshold 10 has been fixed for S because it corresponds to the case of equality of the first bits: the sum of the 4 first b_i (resp. a_i) equals 10 (resp. 11). The dissymmetry between 1 and 0 favours a weak number (4) of synonymy classes of pairs of codons containing C, G, and a large number (8) containing A, U, which explains a more frequent occurrence of AAs coded by triplets with C, G, than AAs coded with A, U (Fig. 2). The threshold 3/32 has been fixed for D to reinforce the similarity with the 2 first bases of a triplet (imposed by the wobble hypothesis).

The Lyapunov function of SUSY is the inverse of the self-similarity 1/S(W,W); it decreases on trajectories until fixed points whose attraction basins are given (boldfaced) by:

$$\begin{aligned} (\mathbf{111111}) \Rightarrow &(111110), (\mathbf{111101}) \Rightarrow (111100), (\mathbf{101111}, \mathbf{101110}, \mathbf{101101}) \\ &\Rightarrow (101100), \end{aligned}$$

$(\mathbf{011111}) \Rightarrow (011110), (\mathbf{011101}) \Rightarrow (011100), (\mathbf{001111}, \mathbf{001110}, \mathbf{001101}) \Rightarrow (001100),$

$(\mathbf{111011}, \mathbf{111010}, \mathbf{111001}) \Rightarrow (111000), (\mathbf{101011}, \mathbf{101010}, \mathbf{101001}) \Rightarrow (101000),$

$(\mathbf{011011}, \mathbf{011010}, \mathbf{011001}) \Rightarrow (011000), (\mathbf{001011}, \mathbf{001010}, \mathbf{001001}) \Rightarrow (001000),$

$(\mathbf{110111}) \Rightarrow (110110), (\mathbf{110101}) \Rightarrow (110100), (\mathbf{100111}) \Rightarrow (100110), (\mathbf{100101}) \Rightarrow (100100),$

$(\mathbf{010111}) \Rightarrow (010110), (\mathbf{010101}) \Rightarrow (010100), (\mathbf{000111}) \Rightarrow (000110), (\mathbf{000101}) \Rightarrow (000100),$

$(\mathbf{110011}) \Rightarrow (110010), (\mathbf{110001}) \Rightarrow (110000), (\mathbf{100011}, \mathbf{100010}, \mathbf{100001}) \Rightarrow (100000),$

$(\mathbf{010011}) \Rightarrow (010010), (\mathbf{010001}) \Rightarrow (010000), (\mathbf{000011}, \mathbf{000010}, \mathbf{000001}) \Rightarrow (000000).$

These basins can be identified in the genetic code, giving the partition of the Fig. 3. Except for the triplet 111100 having a similarity S = 10 with the triplet 101100, the only triplets having a bad assignation are (cf. also Magini and Hornos 2003; Hornos et al. 2004):

011101, which must be related to the block (011111,011110) (S = 9),

110001, which must be related to the block (110101,110100) (S = 9),

010000, whose block has to be related to the block of 100000 (S = 8),

010011, whose block has to be related to the block of 111011 (S = 7).

Coding nucleic bases with three bits, with a similarity S increased of 1 in case of identity of the third bit for the two first bases, would correctly assign the two first

	U	C	A	G
U	UUU Phénylalanine	UCU Sérine	UAU Tyrosine	UGU Cystéine
	UUC Phénylalanine	UCC Sérine	UAC Tyrosine	UGC Cystéine
	UUA Leucine	UCA Sérine	UAA Stop	UGA Stop
	UUG Leucine	UCG Sérine	UAG Stop	UGG Tryptophane
C	CUU Leucine	CCU Proline	CAU Histidine	CGU Arginine
	CUC Leucine	CCC Proline	CAC Histidine	CGC Arginine
	CUA Leucine	CCA Proline	CAA Glutamine	CGA Arginine
	CUG Leucine	CCG Proline	CAG Glutamine	CGG Arginine
A	AUU Isoleucine	ACU Thréonine	AAU Asparagine	AGU Sérine
	AUC Isoleucine	ACC Thréonine	AAC Asparagine	AGC Sérine
	AUA Isoleucine	ACA Thréonine	AAA Lysine	AGA Arginine
	AUG Méthionine/Start	ACG Thréonine	AAG Lysine	AGG Arginine
G	GUU Valine	GCU Alanine	GAU Aspartate	GGU Glycine
	GUC Valine	GCC Alanine	GAC Aspartate	GGC Glycine
	GUA Valine	GCA Alanine	GAA Glutamate	GGA Glycine
	GUG Valine	GCG Alanine	GAG Glutamate	GGG Glycine

Fig. 3 Identification of SUSY attraction basins in the genetic code

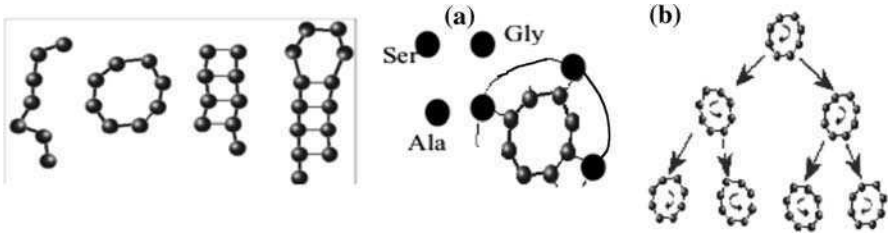


Fig. 4 Primitive forms of XNA (left), linked to AAs (**a**) and replicative dynamics (**b**)

triplets above and relate 010001001 to the block of 101001001 (S = 9) and 010001110 to the block of 110101110 (S = 9). In this case, we would practically always follow the critical similarity threshold equal to 10, by following also the stereochemical theory and the regularities (like in the Gray code) already observed (Swanson 1984; Bosnacki et al. 2003; He et al. 2004; Pohlmeyer 2007).

2.2 A Crucial Variational Principle

Triplets made of nucleic bases (small balls, Fig. 4) are linked to AAs (black ovals), following the genetic code assignation table. These bonds occur in the primitive "soup": in an atmosphere made of atoms of oxygen, hydrogen, nitrogen and carbon, electric discharges (thunderstorms) and ultraviolet or gamma irradiation cause the formation of nucleic bases and amino-acids probably in the same order of apparition than in the Miller's experiment: glycin, alanin, valin, ... (Table 1). The rain water drains these elements toward declivitous parts of the earth relief and evaporation facilitates their confinement in impermeable clays, a kind of "pizza" made of layers of **s**ilicate sandwiching a **g**ibbsite layer in between, in an **s-g-s** stacking sequence. Variable amounts of water, XNA and AA molecules would lie between the s-g-s sandwiches favouring the apparition of the first peptides catalyzed by the primitive XNAs. The primitive forms of XNA are chains, loops or hair-pins. They can be denaturated by numerous physico-chemical factors (cosmic rays, temperature, pH,

Table 1 Amino-Acid frequencies in the human genome (Yarus 2000) and in the genetic code, and AAs apparition ranks in a consensus chronology and Miller's experiment (Oliva et al. 2006)

Amino-Acid	Frequency $\times 10^3$ (human)	Codon frequency $\times 10^3$	Consensus	Miller
Leu	99.1	94	8	8
Ser	79.6	94	7	7
Ala	71.0	63	2	2
Glu	70.1	31	5	5
Gly	66.6	62.5	1	1
Val	61.5	62.5	3	3
Pro	61.1	62.5	6	6
Lys	57.0	31	12	–
Arg	56.8	94	10	–

pressure, hydratation changes, ...). In order to survive during long periods and in such extreme conditions, XNA must be of small size and have to be protected by a molecular hydrophobic shield coming from its direct neighbourhood: AAs are good candidates to do it, but XNAs have to be of sufficient size to offer at least one binding site to each of the AAs present in this neighbourhood. The dual constraint of being of minimal size to escape the denaturation and, at the same time, of maximal size so as the XNA would be protected by AAs constitutes a crucial variational problem whose solution is presented in the following Section.

3 A Primitive ^{t}RNA: The Archetypal Loop AL

This solution consists in a nucleic ring, having (with overlap minimizing its size) one and only one triplet from the 20 AAs synonymy classes like in the primitive «diamond» code of Gamow (Gamow 1954; Hayes 1998). Such a loop does not exist for loops of 20 bases, but the combinatorics exhibits a set called ***A*** (for Archetypal) of solutions consisting in loops of 22 bases. By adding the constraint to have an end codon (simulations show that a preferred breaking point in rings may encourage their replication), AUG is obtained twice in the majority of rings, and in a further selection described below, rings starting with AUG and ending with an end codon become spontaneously the most abundant (Demongeot and Besson 1996; Moreira 2003; Demongeot and Moreira 2007).

3.1 A Plausible Circular or Hairpin-Shaped XNA

We can prove more, that one solution (called «Ancestral Loop» or AL) of the variational problem can be distinguished by using plausibility arguments (Moreira 2003): the loop AL is the barycentre of a selected subset of solutions (made of the solutions having the most stable hairpins as possible secondary structure) for 2 distances and a semi-distance in the space of classes of equivalence of chains for circular permutations: the circular Hamming distance, the distance equal to 22 minus the length of the maximal common substring, and the minimal evolution shuffling similarity, i.e. the minimum number of consecutive deletions of maximal common substrings to do for obtaining the same final sequence. AL is also in mean the closest sequence—for a "cut" distance, corresponding to minimal Hamming distance between four AL and RNA fragments—to about 7,000 transfer RNAs (^{t}RNA) taken from about 180 species (found in literature and databases[1]); reduced to sequences of conserved domains in primary structure, they are considered as relics, invariant between ^{t}RNAs corresponding to different AAs in different species (Hartman 1984). AL is a good candidate of plausible primitive ^{t}RNA (Fig. 5).

The present function of ^{t}RNAs is to link specifically its associated AA in order to build a proteic chain in ribosomes. A primitive version of this function could be done

[1] http://felix.unife.it/Root/d-Biology/d-Genetics-and-evolution/d-tRNA-sequences/t-tRNA-compilation, *Compilation of tRNA and tRNA Gene Sequences*, 1993.

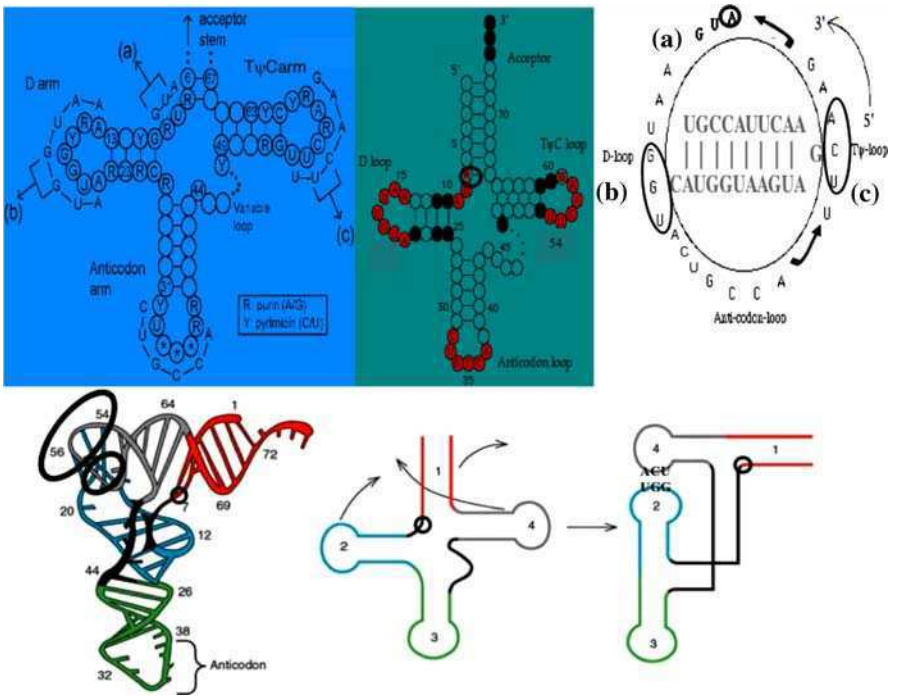


Fig. 5 Plausible primitive ^{t}RNA AL in ring and hairpin secondary structures (top right); ^{t}RNA with invariant loops coming from Gly-^{t}RNA of *Arabidopsis thaliana* (top centre); correspondence with AL (top left); ^{t}RNA ternary structure (bottom left) with base 7 (A) as pivot (bottom centre) in an identical position as on AL between the dual ACU and UGG triplets responsible for the loops 2 and 4 association (bottom right)

by XNA rings like AL, as proposed yet for "aptamers" (Knight and Landweber 1998). The ternary structure of the ^{t}RNAs (Fig. 5) shows 3 critical triplets which serve as mutagenic zones for changing their specificity (Wang and Schultz 2005): the two first triplets (sites 17-18-19) and (54-55-56) are linked, the G19-C56 bond being the most stable (Oliva et al. 2006), and the third one centred on the base A in site 7, serves as pivot for the two orthogonal parts of the tertiary ^{t}RNA structure. These triplets also exist in AL in similar positions. The present ^{t}RNA structure could be obtained through an increasing complexification of the AL kernel: in Fig. 6 the primary sequence of Gly-^{t}RNA of *Arabidopsis thaliana* (see also *Oenethera lamarckiana* in He et al. (2004)) presents 4 parts corresponding to 4 segments of AL and of a virtual ^{t}RNA obtained by coalescence from a RNA ring of length 72 containing all the 64 triplets. As for AL hairpin, this ring could have taken its cloverleaf structure in primitive cells, keeping a specific function in peptide building.

The AL loop contains the 16 pairs of the wobble except CG, the least used pair in the present genomes. The archetypal set ***A*** uses twice the pair UG, the most frequent pair in the chromatin (Trifonov and Sussman 1980), as well as the most frequent codons in numerous animal and vegetal species (Demongeot and Besson 1996; Moreira 2003) and presents without overlap 2 series of 7 contiguous triplets. We

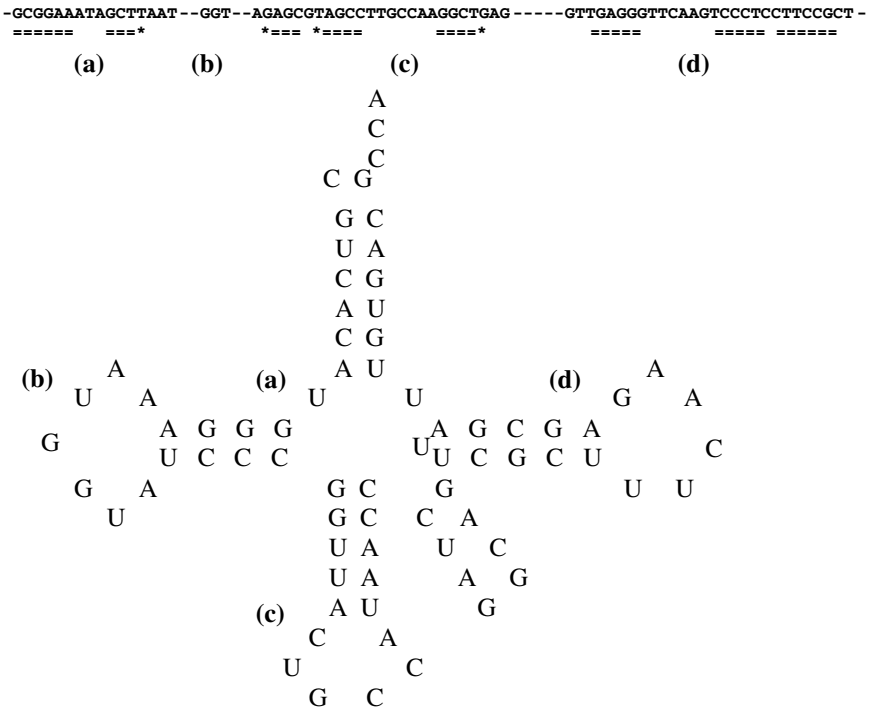


Fig. 6 Primary sequence of Gly-^{t}RNA of *Oenothera lamarckiana* (top) and RNA ring of length 72 in cloverleaf secondary structure (bottom)

indicate into parentheses the rank of the triplets in the order of increasing molecular complexity of the corresponding 12 AAs not start nor stop (Trifonov 2000), close to their rank of occurrence in Miller's experiment (Table 1):

GGU (1) ACU (9) GCC (2) AUU (10) CAA (14) GAU (4) GAA (5) Total: 45,
UCA (7) AGA (12) UGA AUG GUA (3) CUG (8) CCA (6) Total: 36

The sum of these ranks equals 81, close to the sum of the 12 first integers (78), corresponding to the case of the use in AL of the minimal ranked AAs.

4 AL And Present RNA Relics

We can now report several properties of present RNAs in adequation with the AL structure:

(1) the XNA secondary structures the most frequently proposed as common ancestors for the present ^{t}RNAs are rings or hairpins (di Giulio 1992, 1997; Hopfield 1978; Toulokhonov et al. 2001; Poole et al. 1998; Shimizu 1995; Eigen et al. 1981; Szathmary and Maynard Smith 1997). The most invariant

parts of the present ^{t}RNAs are their loops (Szathmary and Maynard Smith 1997; Rodin et al. 1993), whose sequences after cancelling their stems, are close to AL, which is then a candidate for being a ^{t}RNA ancestor acting as an aptamer, i.e. a primordial ^{t}RNA as advanced by Yarus (1989) or serving as a template for building the first peptides, these ones catalyzing later the XNA synthesis in a positive evolutionary loop. The argument "extented anti-codon" claims that the ^{t}RNA–AA bond can involve bases of the anti-codon loop other than strictly those of the anti-codon (Majerfeld et al. 2005; Knight and Landweber 1998), this mechanism being similar to an aptameric AL-AA link (especially with amino-acids like arginine, lysine or tyrosine).

(2) AL is solution of the original combinatorial problem (Demongeot 1978) and is close to micro-RNAs (Zhang et al. 2005) considered as interspecific relics present in the non-coding DNA and able after maturation to hybridize mRNAs, to inhibit the proteic translation (Fig. 7). Hairpin form of AL is

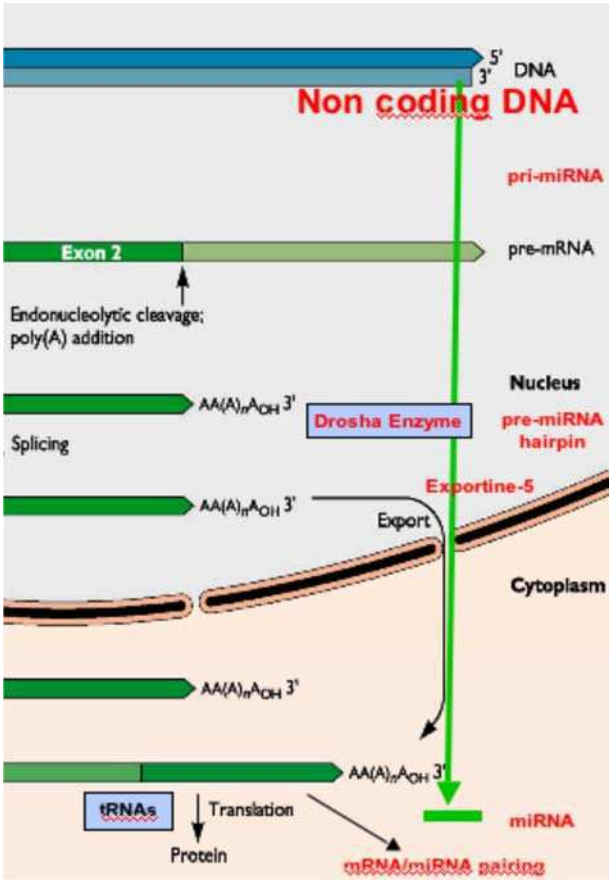
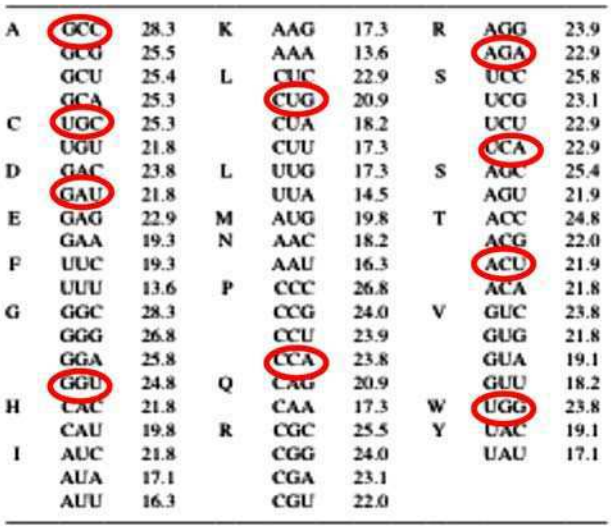


Fig. 7 Micro-RNAs maturation and control of translation



A	GCC	28.3	K	AAG	17.3	R	AGG	23.9
	GCG	25.5		AAA	13.6		AGA	22.9
	GCU	25.4	L	CUC	22.9	S	UCC	25.8
	GCA	25.3		CUG	20.9		UCG	23.1
C	UGC	25.3		CUA	18.2		UCU	22.9
	UGU	21.8		CUU	17.3		UCA	22.9
D	GAC	23.8	L	UUG	17.3	S	AGC	25.4
	GAU	21.8		UUA	14.5		AGU	21.9
E	GAG	22.9	M	AUG	19.8	T	ACC	24.8
	GAA	19.3	N	AAC	18.2		ACG	22.0
F	UUC	19.3		AAU	16.3		ACU	21.9
	UUU	13.6	P	CCC	26.8		ACA	21.8
G	GGC	28.3		CCG	24.0	V	GUC	23.8
	GGG	26.8		CCU	23.9		GUG	21.8
	GGA	25.8		CCA	23.8		GUA	19.1
	GGU	24.8	Q	CAG	20.9		GUU	18.2
H	CAC	21.8		CAA	17.3	W	UGG	23.8
	CAU	19.8	R	CGC	25.5	Y	UAC	19.1
I	AUC	21.8		CGG	24.0		UAU	17.1
	AUA	17.1		CGA	23.1			
	AUU	16.3		CGU	22.0			

ugaa(g)gg ugc miR 319 Saccharum oficinarum ; **aauggu(u) cc(c)uu(u)a** hsa miR 522

Fig. 8 Triplets thermo-denaturation energies in kcal/M (top) (Trifonov 2000) and matching sequences between two AL and vegetal (Gottesman 2005) or human (Bentwich et al. 2005) micro-RNAs (bottom)

thermodynamically stable[2], as shown by the triplet thermo-denaturation energies: stable AL triplets encircled on Fig. 8 (Wang et al. 2006; Trifonov 2000) form a structure, local optimum of thermostability for the Biopolymer Chain Elasticity Algorithm (Cognet 2006) and sub-optimum for simple models of nearest neighbours (Freier et al. 1986; Dale et al. 2000), the sub-optimality being better because of the double constraint to be hairpin stable as well as easy to reconfigure into a ring to ensure the primitive peptide building function.

(3) The small RNAs (sRNAs) comprise the micro-RNAs and the small interfering RNAs (siRNAs). Both can inhibit the translation (notably the micro-RNAs) as well as the replication of some viral genomes. For example, they can interfere with the expression of viral genomes due to a selective hybridization with target viral mRNAs, as for the human HDV virus (Chang and Taylor 2003); HDV genome shows good matching scores between its target sequences and anti-AL (average match 9.63/21) or AL (average match 9.5/21) for the circular Hamming distance (Fig. 9), calculated by counting the maximal number of matches with a circular permutation of AL (matching score) and then substracting this score to the length of the sequence.

We can also calculate (Fig. 10) the number of matching segments coming from various micro-RNAs and covering a given position of the untranslated (5′ UTR) HCV viral genome, related to the number of microRNAs in each studied species.

[2] http://helix.nih.gov/docs/online/mfold/node5.html

siRNA targets (HDV genome)	# ALmatches	phase
AAGAAAGAAGUUAGAGGAACU	11	1
AAGAUAGAGGACGAAAAUCCC	9	8
AACGGACCAGAUGGAGGUAGA	9	8
AAGGAAGGCCCUCGAGAACAA	10	10
AACAAGAAGAAGCAGCUAUCG	9	4
AAGAACCUCAGCAAGGAGGAA	9	13
AAGAGGAACUCAGGAGGUUGA	9	11
AAGACGAGAGAAGGGAAAGAA	8	4
AAACCAGGGAUUUCCAUAGGA	10	4
AAAGAGCAUUGGAACGUCGGA	11	2
AAGGGUUGAGUAGCACUCAGA	10	3
AAGCGAGGAGGAAAGCAAAGA	9	5
AACUCGACUUAUCGUCCCCAU	11	1
AAUGCUCUUUACCGUGACAUC	10	11
AAGCGCCUCUUGUUCGCUGAA	8	9
AAGUCGAGUUCCCCGGGAUAA	9	6

AAGAUGAAUGGUACUGCCAUUC

Fig. 9 Human siRNA 5′-3′ targets in the HDV viral genome (left); matching score with AL and matching phase (right). AL sequence is indicated at the bottom

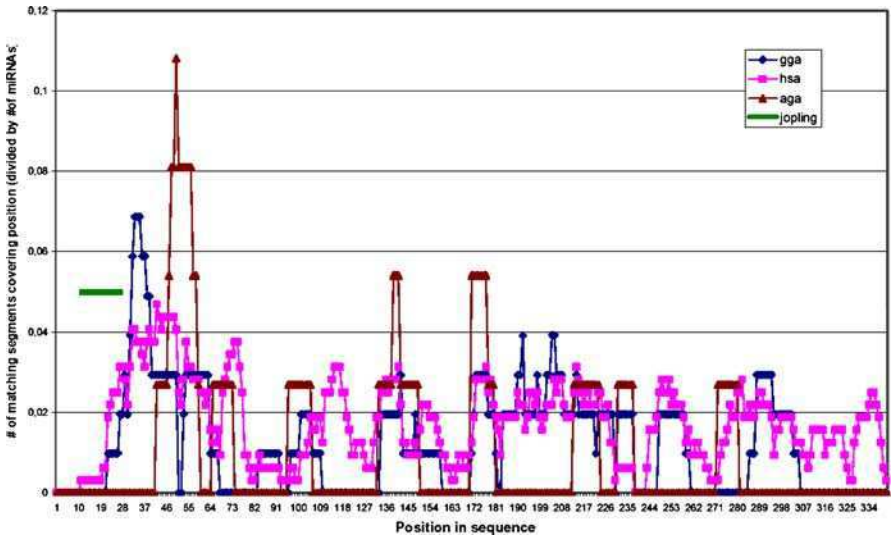


Fig. 10 Matching between a sequence of the 5'UTR HCV viral genome and micro-RNAs from *Gallus gallus* (gga), *Homo sapiens* (hsa) and *Anopheles gambiae* (aga). The position of the human miR 122 is indicated in green on the beginning of the viral genome

These calculations show a better match for host species (*Gallus gallus* and *homo sapiens*) than for the vector (*Anopheles gambiae*). A particular micro-RNA, the human miR 122 (Jopling et al. 2005) has for example an excellent fit between its sequence ACACCATTGTCACACTCCA (located between the positions 5 and 23

of miR 122) and the complementary of the HCV sequence ACACACTAGG TACACTCCA (located between the positions 7 and 25 of the 5′ UTR HCV genome). More generally, we can study the variability of the matching scores (for the circular Hamming distance) by calculating the average matches between presently known micro-RNAs repeated in at least two different species (repmirs) and random small RNAs of length 22 having the same base frequencies than AL. The average matches between repmirs and AL and between siRNA targets (from Fig. 9) and AL are indicated on Fig. 11 on the right part of the graph of the distribution function: they are significantly better than the random matching scores.

Finally, the whole set of the random sRNAs of length 22 has a barycenter farther from t, a, s (respectively barycenters of the 3 subsets: set of the ^{t}RNA loops ordered in the primary order, set $\boldsymbol{A}$ of the ancestral rings, solutions of the variational problem above, and set of the real small RNAs, made of siRNAs and micro-RNAs), than t, a, s between themselves (Figs. 12, 13). The calculations have been made (Moreira 2003; Demongeot and Moreira 2007) for the "cut" distance defined above and for the circular version of the classical edit distance. We can notice than the mean length of the presently known human micro-RNAs is 22.

The results presented above show the proximity (by comparison with random sequences) between the ancestral ring AL and the present RNA relics like the invariant parts of the present tRNAS or the small RNAs, reinforcing the idea of a common ancestor for these functionally very important RNA structures.

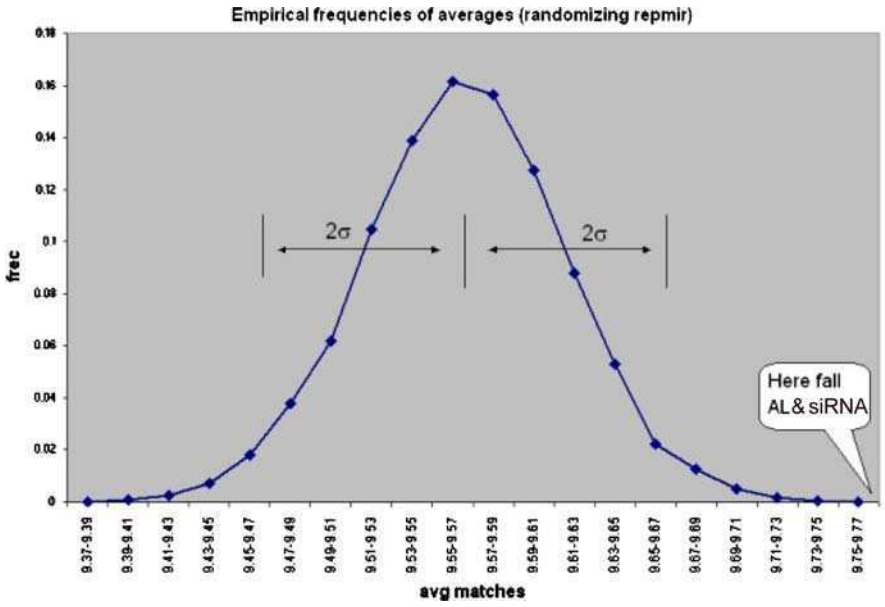


Fig. 11 Histogram of match averages between micro-RNAs repeated in at least 2 different species (repmirs) and random small RNAs of length 22 having the same base frequencies than AL. The average matches between the repmirs and AL, and between HDV siRNA targets and AL are indicated at the right part of the curve

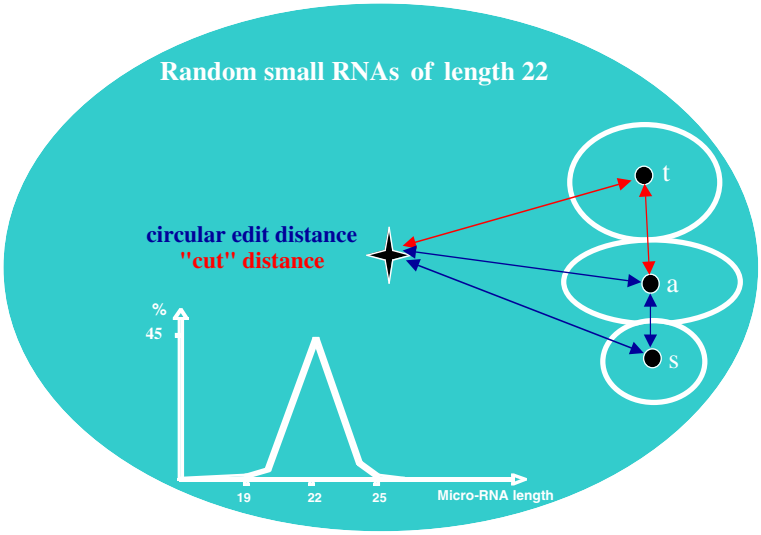


Fig. 12 The set of the random small RNAs of length equal to the mean (22) of the lengths of 319 human micro-RNAs (histogram bottom), with indication of the barycenters of the ^{t}RNA loops set (t), ancestral RNA rings set ***A*** (a) and real small RNAs set (s)

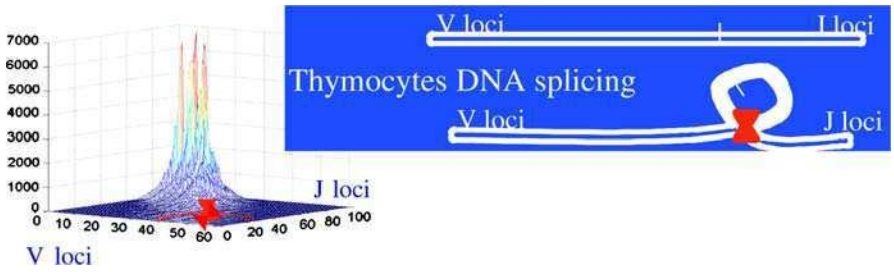


Fig. 13 Thymocytes DNA splicing giving an evolving neo-genome creating neo-genes from homologous (proximal and distal) V and J loci (after Pasqual et al. 2002)

5 Immune Genome

Immune genome appeared before chordates and gave initially birth to a non-specific repertoire which did not recognize exactly the endogeneous proteome. This primitive immune system expressed anti-proteins, and then anti-anti-proteins identical to the initial ones concerning catalytic and regulatory sites but evolving to minimal architectures, keeping only parts necessary to their function, playing in a new deal between genome and proteome. During this first phase we can imagine that not only immune proteins could interact with the infectious ones, but also micro-RNAs, able to hybridize with mRNAs or interact with peptides of viral or bacterial agents and present micro-RNAs could be relics of this ancestral immunity (Berezikov and Plasterk 2005). The immune antiviral response of the micro-RNAs has been recently shown (Jopling et al. 2005; Lecellier et al. 2005).

The neo-genome produced in human by chromosome 14 is richer than the non-immune genome (about 100,000 potential anti-body neo-genes against about 25,000 non-immune genes, cf. Baum et al. 2004). We can conceive the immune machinery as able to evoke ancient aggressions favouring the recombination of fossil anti-body genes. As shown above, numerous sequences of viral and retro-viral (Buzdin et al. 2003; Hughes and Coffin 2004) genomes are present in human introns as well as their "anti-genes" coding for the anti-bodies of their translated proteins.

6 Intraspecific Diversity and Interspecific Divergence

Physiologic rearrangements (crossing-overs) as well as pathologic ones (translocations, deletions, inversions, insertions) insure the intraspecific diversity into species having a sexuality, through the segregation of loci controlling their phenotypic characteristics. As well as punctual mutations due to environmental physico-chemical factors affecting the genome of any organism, constitutional abnormalities like chromosomal translocations can contribute to the intraspecific divergence. On Fig. 14, we can see a correlation between the localizations of the rearrangements due to translocations or crossing-overs on the human chromosome 3, and those of the genic expression, in particular the ubiquitary one which occurs during all the cell cycle (Faraut and Demongeot 2000; Demongeot et al. 2000). This correlation is not surprizing because that means that DNA sites being in transcription are more fragile than those that are compacted in the nuclear chromatin. Therefore, this observation

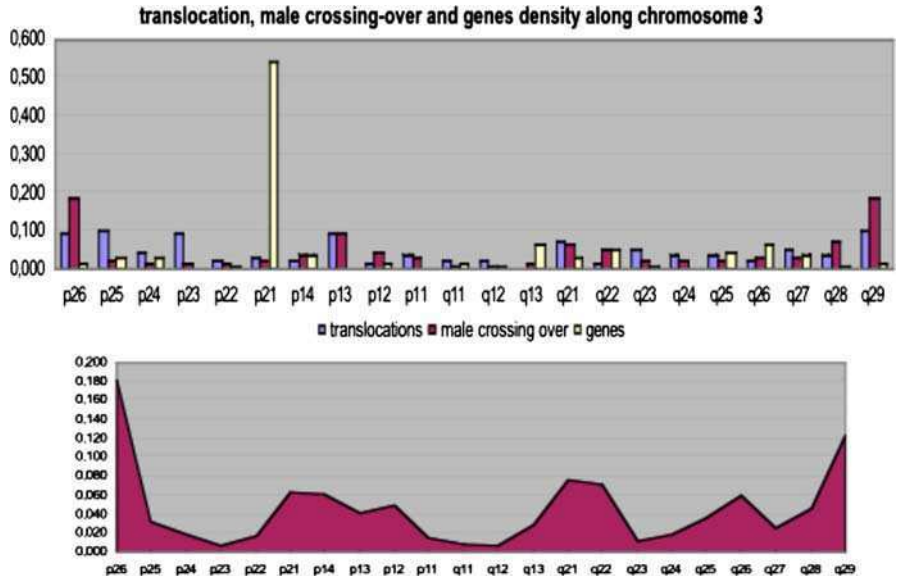


Fig. 14 Correlation between hot spots of rearrangements both physiologic (crossing-overs) and pathologic (translocations), and genic expression modes (for all genes on the top histogram and for ubiquitary genes on the bottom) along the human chromosome 3

has big consequences because it relates the evolution of the genes to their expression possibly linked to the environment (for example, genes expressed in aerobic conditions are not the same than those expressed in anaerobic ones); hence the constitutional chromosomic abnormalities causing the interspecific divergence could depend on environmental factors.

It is now clear that the interaction between nucleic and AAs constitutes the essence of life, as we have shown for the first steps of the life's start. There is no life if the first assemblages occur in isolated XNA or AA worlds and there is life when these worlds cooperate. The evolution of the present life forms could follow two scenarios among several others: (i) proteins could evolve toward an autonomic functioning, their reproduction using a proteic template, like in prion case (Laurent 1996) and species could disappear by accumulating a "parasitic" protein taking the place of functional ones. The extinction of the present life forms could occur by dominance of this protein if its malthusian parameter (growth rate) is larger than those of the other proteins, and ii) a similar evolution does not concern the nucleic acids, because they need a proteic environment for ensuring a fast and directed replication, and the co-evolution of both nucleic and proteic populations would prevent their segregation, but only favour the possible occurrence of an XNA more thermodynamically stable. We can then represent in a synthetic way (Fig. 15) the possible successive steps of the life evolution, from its start until its possible end or next step, each being speculative, but plausible. This plausibility (and its falsifiability) is based on direct affinities between AAs and their codons and anti-codons (Pelc and Welton 1966; Hendry et al. 1981; Hobish et al. 1995; Yarus 2000;

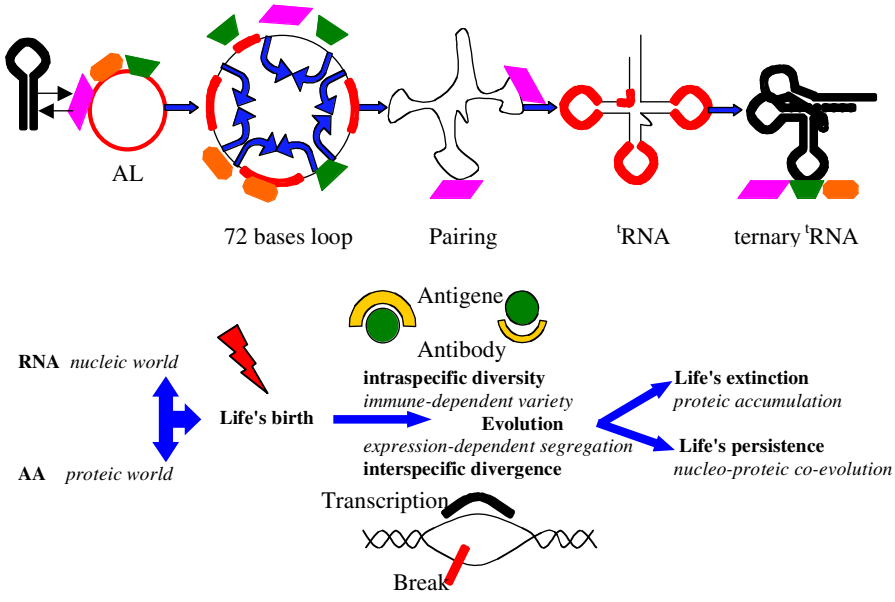


Fig. 15 Plausible scheme of the life's evolution, with the different forms of tRNAs, archetypal and present (top), and protein diversity and divergence (bottom)

Majerfeld et al. 2005), which can justify the existence of a minimal ancestral ring AL as well as rings or hairpins (such as aptamers), like those of Fig. 6 with codons having the same frequencies than in genetic code and than in the most frequent AAs represented in the human genome and in the Miller's experiment (Table 1).

7 Optimality of the Genetic Code

The mutual benefit the XNA and AA worlds have found from their direct association (e.g. Leu, Glu, Val et Arg preferentially linked to their codons and anti-codons, within aptamers (Knight and Landweber 1998; Sciarrino 2003)) is compatible with their co-evolution under constraints well summarized in de Duve (2002): «The theory considered most likely today supposes a historical, co-evolutionary process in which the anti-codons and the corresponding amino-acids were progressively recruited together under the control of natural selection. Several arguments support this hypothesis. The most convincing lies in the structure of the code, which, far from being random, happens to be such as to minimize the deleterious consequences of mutations.».

We can express the genetic code adaptation by using a variational criterion based on the dual principle of minimizing its mutation function M and maximizing its information I. Let us take the example of a genetic code with only 2 Aas.

Figures 16 and 17 above gives the optimal frequency $f_o \approx 2/3$ for a 2 Aas code, obtained at the intersection of the graphs of $M(f) = 2f(1-f)$ and $I(f) = (-f\mathrm{Log}f - (1-f)\mathrm{Log}(1-f))\int_0^1 M(x)dx / \int_0^1 H(x)dx$, where $H(x) = -x\mathrm{Log}x - (1-x)\mathrm{Log}(1-x)$. If the frequency of the amino-acid AA_i equals $n_i/64$, where n_i is the size of its synonymy class, the mutation function M is obtained by dividing by 64 the

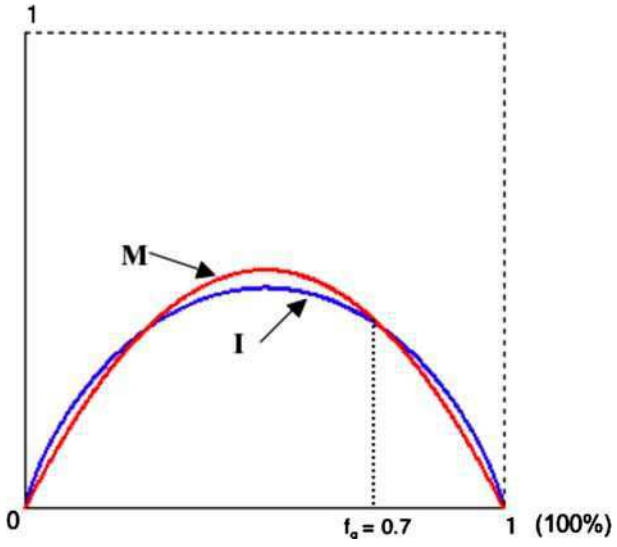


Fig. 16 Graphs of the mutation function M and of the information I

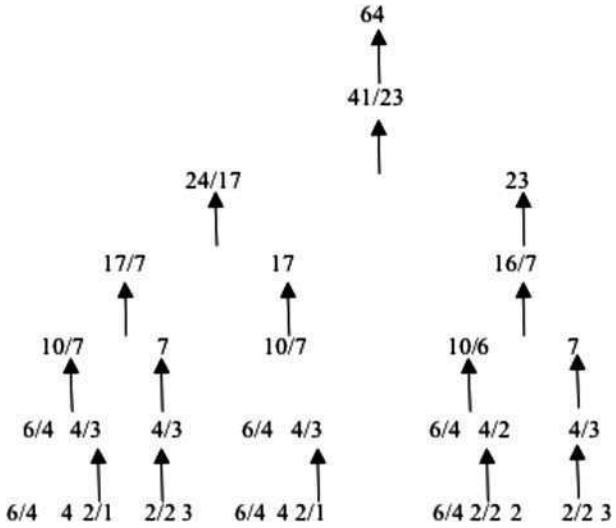


Fig. 17 Renormalization tree of synonymy classes of the genetic code, respecting the optimal proportion f_o

deleterious (i.e. provoking the exit from the class) mutations expectation E equal to $E = \sum_{i=1,2} (64 - n_i)n_i/64$.The function I is just the entropy of the distribution of the AA frequencies, normalized by a quantity ensuring for M and I the same mean value on [0,1].

If now we renormalize the dispatching of the genetic code into 3 classes of 6 codons, 5 of 4, 2 of 3, 9 of 2, and 2 of 1, following at each renormalization step the rule which consists in roughly respecting the optimal frequencies ($f_o \approx 2/3$, $1 - f_o \approx 1/3$) by dividing the classes to renormalize. Then we obtain a coherent renormalization tree verifying at each node the variational criterion above. In the reverse sense, this tree can be described by first combining the classes of size 1 and 2 (bottom line of the tree) for obtaining classes of size 3 and 4, then combined to classes of size 4 and 6 for obtaining an assignment of the 64 codons to 8 AAs (e.g. the first 8 AAs of the Miller's experiment). This condensed proof explains the classical property of resistance to the mutations of the genetic code (Labouygues 1976; Figureau and Pouzet 1984), compatible with its congruence to the physico-chemical properties of AAs shown above by the automaton SUSY.

The genetic code can be then considered as produced both by initial conditions and by evolution of the life dynamics: it represents an optimal compromise between resistance to mutations, maximal information and adequation to the binding potential to AAs.

8 Conclusion

Life is born from the meeting between two entities, nucleic acids and amino acids, each of them being able to independently create polymers, but taking advantage

from the ones to the others for accelerating their multiplication and survival processes, thus offering a more important choice to the selective mechanisms. Fundamental steps like the refinement of the proteic function and the rearrangement of the genomic memory then permitted the evolution of species, with the intraspecific diversity and the interspecific divergence described by the cladistics, and showing an increasing complexity of the nucleic and proteic populations of molecules in interaction. This evolution could lead to the extinction of the present life forms, by dominance of a population (e.g. the proteic one winning by autocatalysis and accumulation, process symmetrical of the starting game, but without XNA), or to a continuation of life in a co-evolution of the two entities through a multi-level organization (until the social structure) increasing in the same evolutionary process its complexity and auto-protective ability against the environmental degradation.

Acknowledgements We are very indebted to C. de Duve, M. Eigen, J. Maynard-Smith (died in 2004), A. Pacault and J. Tonnelat (both died in 2008), for their marvellous books, and also to F. Kepes for fruitful discussions.

References

Atlan H (1972) L'organisation Biologique et la Théorie de l'Information. Hermann, Paris

Atlan H (2004) Etincelles de Hasard. Le Seuil, Paris

Baum TP, Pasqual N, Thuderoz F, Hierle V, Chaume D, Lefranc MP, Jouvin-Marche E, Marche P, Demongeot J (2004) IMGT/GeneInfo: enhancing V(D)J recombination database accessibility. Nucleic Acids Res 32:51–54

Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, Barzilai A, Einat P, Einav U, Meiri E, Sharon E, Spector Y, Bentwich Z (2005) Identification of hundreds of conserved and non conserved human μRNAs. Nat Genet 17:766–770

Berezikov E, Plasterk RHA (2005) Camels and zebrafish, viruses and cancer: a microRNA update. Hum Mol Genet 14:183–190

Blalock JE, Bost K (1986) Binding of peptides that are specified by complementary RNAs. Biochem J 234:679–683

Bosnacki D, ten Eikelder H, Hilbers P (2003) Genetic code as a Gray code revisited. In: Proceedings of the international conference on mathematics and engineering techniques in medicine and biological sciences METMBS'03. CSREA Press, Athens, Georgia, pp 447–456

Buzdin A, Gogvadze E, Kovalskaya E, Volchkov P, Ustyogova S, Illarionova A, Fushan A, Vinogradova T, Sverdlov E (2003) The human genome contains many types of chimeric retrogenes generated through in vivo recombination. Nucleic Acids Res 31:4385–4390

Chang J, Taylor JM (2003) Susceptibility of human hepatitis delta virus RNAs to small interfering RNA action. J Virol 77:9728–9731

Cognet J (2006) Personal communication

Dale T, Smith R, Serra MJ (2000) A test of the model to predict unusually stable RNA hairpin loop stability. RNA 6:608–615

de Duve C (2002) Life evolving. Oxford University Press, Oxford UK

Demongeot J (1978) Sur la possibilité de considérer le code génétique comme un code à enchaînement. Revue de Biomaths 62:61–66

Demongeot J, Besson J (1983) Code génétique et codes à enchaînement I. C R Acad Sci 296:807–810

Demongeot J, Besson J (1996) Genetic code and cyclic codes II. C R Acad Sci 319:520–528

Demongeot J, Moreira A (2007) A possible circular RNA at the origin of life. J Theor Biol 249:314–324

Demongeot J, Aracena J, Ben Lamine S, Mermet MA, Cohen O (2000) Hot spots in chromosomal breakage: from description to etiology. In: Sankoff D, Nadeau JH (eds) Comparative genomics. Kluwer, Amsterdam, pp 71–85

Demongeot J, Elena A, Weil G (2006) Potential automata. Application to the genetic code III. C R Biol 329:953–962
di Giulio M (1992) On the origin of the tRNA molecule. J Theor Biol 159:199–214
di Giulio M (1997) On the origin of the genetic code. J Theor Biol 187:573–581
Eigen M (1971) Molekuläre selbstorganisation und evolution. Naturwissenschaften 58:465–523
Eigen M, Gardiner W, Schuster P, Winkler-Oswatitsch R (1981) The origin of genetic information. Sci Am 244:88–92
Faraut T, Demongeot J (2000) Benefits of a model of segregation for the understanding of chromosomal evolution. In: Sankoff D, Nadeau JH (eds) Comparative genomics. Kluwer, Amsterdam, 13–17
Figureau A, Pouzet M (1984) Genetic code and optimal resistance to the effects of mutations. Orig Life 14:579–588
Freier SM, Kierzek R, Jaeger JA, Sugimoto N, Caruthers MH, Neilson T, Turner DH (1986) Improved free-energy parameters for predictions of RNA duplex stability. Proc Natl Acad Sci USA 83: 9373–9377
Gamow G (1954) Possible relation between deoxyribonucleic acid and protein structures. Nature 173:318–319
Gilis D, Massar S, Cerf N, Rooman M (2001) Optimality of the genetic code with respect to protein stability and amino-acids frequencies. Genome Biol 2:1–12
Glansdorff P, Prigogine I (1971) Structure, stabilité et fluctuations. Masson, Paris
Gottesman S (2005) μ for microbes: noncoding regulatory RNA in bacteria. Trends Genet 21:399–404
Hartman H (1984) Speculations on the evolution of the genetic code III. Orig Life 14:643–648
Hayes B (1998) The invention of the genetic code. Am Sci 86:8–14
He M, Petoukhov S, Ricci P (2004) Genetic code, Hamming distance and stochastic matrices. Bull Math Biol 66:1405–1421
Hendry LB, Bransome ED, Hutson MS, Campbell LK (1981). First approximation of a stereochemical rationale for the genetic code based on the topography and physicochemical properties of 'cavities' constructed from models of DNA. Proc Natl Acad Sci USA 78:7440–7444
Hobish MK, Wickramasinghe NSMD, Ponnamperuma C (1995) Direct interaction between amino-acids and nucleotides as a possible physico-chemical basis for the origin of the genetic code. Adv Space Res 15:365–375
Hopfield J (1978) Origin of the genetic code: a testable hypothesis based on tRNA structure, sequence, and kinetic proofreading. Proc Natl Acad Sci USA 75:4334–4338
Hornos JEM, Braggion L, Magini M, Forger M (2004) Symmetry preservation in the evolution of the genetic code. Life 56:125–130
Hughes J, Coffin J (2004) Human endogeneous retrovirus K solo-LTR formation and insertional polymorphism: implications for human and viral evolution. Proc Natl Acad Sci USA 101: 1668–1672
Jopling CL, Yi M, Lancaster AM, Lemon SM, Sarnow P (2005) Modulation of hepatitis C virus RNA abundance by a liver-specific miR. Science 309:1577–1581
Knight R, Landweber L (1998). Rhyme and reason: RNA-arginine interactions and the genetic code. Chem Biol 5:215–220
Labouygues JM (1976) New mathematical model of genetic code with passive resistance to mutations or buccion and complementary dynamic protective mechanisms against noise at genome level. Agressologie 17:329–335
Laurent M (1996) Prion diseases and the «protein only» hypothesis: a theoretical dynamic study. Biochem J 318:35–39
Lecellier CH, Dunoyer P, Arar K, Lehmann-Che J, Eyquem S, Himber C, Saïb A, Voinnet O (2005) A cellular microRNA mediates antiviral defense in human cells. Science 308:557–560
Magini M, Hornos JEM (2003) A dynamical system for the algebraic approach to the genetic code. Braz J Phys 33:825–830
Majerfeld I., Puthenvedu D, Yarus M (2005) RNA affinity for molecular L-histidine; Genetic code origins. J Mol Evol 61:226–235
Mitaku S, Hirokawa T, Tsuji T (2002) Amphiphilicity index of polar amino-acids as an aid in the characterization of AA preference at membrane-water interfaces. Bioinformatics 18:608–616
Moreira A (2003) Partículas y agentes simples en autómatas celulares y otros sistemas discretos. PhD Thesis, Universitad de Chile, Santiago de Chile
Oliva R, Cavallo L, Tramontano A (2006) Accurate energies of hydrogen bonded nucleic acid base pairs and triplets in tRNA tertiary interactions. Nucleic Acids Res 34:865–879

Oparine A (1924) Proiskhzhdenie Zhzni. Ind Moskovskii Rabochii, Moscow
Pasqual N, Gallagher M, Aude-Garcia C, Loiodice M, Thuderoz F, Demongeot J, Ceredig R, Marche P, Jouvin-Marche E (2002) Quantitative and qualitative changes in ADV-AJ rearrangements during mouse thymocytes differentiation: implication for a limited TCR ALPHA chain repertoire. J Exp Med 196:1163–1174
Pelc SR, Welton MGE (1966) Stereochemical relationship between coding triplets and amino-acids. Nature 209:868–870
Pohlmeyer R (2007) http://www.biology-online.org/biology-forum/about11707.html
Poole A, Jeffares D, Penny D (1998) The path from the RNA world. J Mol Evol 46:1–17
Rodin S, Ohno S, Rodin A (1993) tRNAs with complementary anticodons: could they reflect early evolution of discriminative genetic code adaptors? Proc Natl Acad Sci USA 90:4723–4727
Sciarrino A (2003) A mathematical model accounting for the organization in multiplets of the genetic code. BioSystems 69:1–13
Shimizu M (1995) Specific amino-acetylation of C4N hairpin RNAs with the cognate aminoacyl-adenylates in the presence of a dipeptide: origin of the genetic code. J Biochem (Tokyo) 117:23–26
Swanson R (1984) A unifying concept for the amino-acid code. Bull Math Biol 46:187–203
Szathmary E, Maynard Smith J (1997) From replicators to reproducers: the first major transitions leading to life. J Theor Biol 187:555–571
Thom R (1972) Stabilité structurelle et morphogenèse. Benjamin, New York
Toulokhonov I, Artsimovitch I, Landick R (2001) Allosteric control of RNA polymerase by a site that contacts nascent RNA hairpins. Science 292:730–733
Trifonov E (2000) Consensus temporal order of amino-acids and evolution of the triplet code. Gene 261:139–151
Trifonov E, Sussman J (1980) The pitch of chromatin DNA is reflected in nucleotide sequence. Proc Natl Acad Sci USA 77:3816–3820
Trinquier G, Sanejouand YH (1998) Which effective property of amino-acids is best preserved by the genetic code? Protein Eng 11:153–169
Wang L, Schultz P (2005) Expanding the genetic code. Angew Chem Int Ed 44:34–66
Wang YL, Bao J, Sun Y, Yang J (2006) Energy and structural analysis of double nucleic acid triplets. J Theor Biol 238:85–103
Welton MGE, Pelc SR (1966) Specificity of the stereochemical relationship between ribonucleic acid-triplets and amino-acids. Nature 209:870–872
Yarus M (1989) Specificity of arginine binding by the tetrahymena intron. Biochemistry 28:980–988
Yarus M (2000) RNA-ligand chemistry: a testable source for the genetic code. RNA 6:475–484
Zhang BH, Pan XP, Wang QL, Cobb G, Anderson T (2005) Identification and characterization of new plant μRNAs using EST analysis. Cell Res 15:336–360